Invited Article

AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Otolaryngology–Head and Neck Surgery
144(6) 831–837

DOI: 10.1177/0194599811399724
http://otojournal.org
SAGE

A Practical Guide to Surveys and Questionnaires

Eric L. Slattery, MD[1], Courtney C. J. Voelker, MD, DPhil[1], Brian Nussenbaum, MD[1], Jason T. Rich, MD[1], Randal C. Paniello, MD[1], and J. Gail Neely, MD[1]

No sponsorships or competing interests have been disclosed for this article.

Abstract

Surveys with questionnaires play a vital role in decision and policy making in society. Within medicine, including otolaryngology, surveys with questionnaires may be the only method for gathering data on rare or unusual events. In addition, questionnaires can be developed and validated to be used as outcome measures in clinical trials and other clinical research architecture. Consequently, it is fundamentally important that such tools be properly developed and validated. Just asking questions that have not gone through rigorous design and development may be misleading and unfair at best; at worst, they can result in under- or overtreatment and unnecessary expense. Furthermore, it is important that consumers of the data produced by these instruments understand the principles of questionnaire design to interpret results in an optimal and meaningful way. This article presents a practical guide for understanding the methodologies of survey and questionnaire design, including the concepts of validity and reliability, how surveys are administered and implemented, and, finally, biases and pitfalls of surveys.

Keywords

surveys, questionnaires, survey design, instrument, psychometrics

Received December 19, 2010; revised January 17, 2011; accepted January 18, 2011.

We encounter surveys and questionnaires in almost every aspect of our daily lives. Whether we are shopping in the mall, checking our e-mail, or answering the telephone, someone always seems to be asking our opinion on some issue. Why should we care about the quality of these surveys? The data collected from these questions may be used to make larger societal decisions. Surveys heavily influence politics, help shape public policy decisions, and affect product development.

Surveys and questionnaires have important roles in decision making for medical professionals. For example, examination of quality of life and other patient views toward disease and interventions have taken priority in research, making it imperative that clinicians are able to interpret these data. In addition, surveys with questionnaires may be the only means of broadly evaluating problems and developing corrective actions for rare or unusual events, such as near-miss catastrophic events in surgery. Surveys with questionnaires may be developed to assess symptoms and ultimately used as outcome measures in clinical trials.[1] Therefore, because the results of surveys are used in such powerful ways, survey quality is vital.

The purpose of this article is to present a practical guide for understanding the methodologies of survey and questionnaire design, including the concepts of validity and reliability, how surveys are administered and implemented, and, finally, biases and pitfalls of surveys.

Surveys, Questionnaires, and Psychometrics Defined

Surveys and questionnaires are not synonymous. A survey is a general methodology for gathering, describing, and explaining information from sample(s) to construct a quantitative description of a population.[2,3] Survey research is 1 of the 3 techniques for collection of primary data—the other 2 being direct measurement and observation.[4] Depending on the purpose of the survey, the data can be reported in a variety of forms and can be categorized according to design (eg, prospective or retrospective) or by type of data collected (eg, continuous, categorical, or nominal). Information within surveys can be gathered through many

[1]Practical Guides Writing Group, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, St Louis, MO, USA

Corresponding Author:
J. (John) Gail Neely, MD, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, 660 S Euclid Ave, Box 8115, St Louis, MO 63110, USA
Email: jgneely@aol.com

means, including face-to-face interviews, telephone interviews, or most commonly through self-administered questionnaires.[5]

Questionnaires refer to a specific tool, also known as an *instrument*, for gathering information. Questionnaires are also known as *scales* when their assessment creates a quantified score.[6] Questionnaires consist of a series of questions and are usually self-administered. The questions contain specific concepts of interest or *items* deemed worthy of investigation and can be disseminated in a variety of ways, including mail, Internet, or even read to participants.[3] The rest of this article concentrates on questionnaires.

Critical to obtaining accurate information is using a well-constructed instrument. The field of psychometrics (measuring psychological behaviors or responses) involves the study and creation of psychological tests, including questionnaires.[7] It includes the selection of items (questions), pilot testing the questions, recognizing bias, and factor analysis. The fundamentals of validity and reliability derive from psychometrics.[8] The application of psychometrics ultimately strengthens instruments within surveys and the conclusions drawn from them.

Survey and Questionnaire Design

When planning a survey, it may not be necessary to develop an instrument de novo. Many validated instruments already exist and can often be used directly or adapted for various uses (example online sources to find validated instruments: www.nlm.nih.gov/nichsr/hsrr_search and http://outcomes-trust.org/index.html). It should be noted that if a previously validated instrument is used in an unintended way, the instrument becomes invalidated. However, despite a growing number of questionnaires and scales, new ones may be needed to gather specific information. Various design theories exist, including classic test, generalizability, and item response theories. Concepts of validation and reliability derive from classic test theory. The details of these theories are beyond the scope of this article but may be seen in the review by Streiner and Norman.[6]

Determine the Objective

The first step of survey construction is agreeing on an objective (**Table 1**). Study design, structure, type of data, and analysis will ultimately depend on a given survey's objective. For example, if a comparison is attempted with a group prior to and after an intervention, a prospective study is usually created. Often the questionnaire used will be a scale allowing for quantification. Analysis of data depends on study design, number of groups, and type of data (eg, nominal, ordinal, or continuous).[9]

Designing Questions

After defining the objective, the next step is deciding how to best obtain the information. If no suitable questionnaire exists for a given objective, then one will need to be created. In designing an instrument, specific items must be agreed upon by experts in the field and those with the problem as important for further evaluation. The research subjects, which in health-related questionnaires often are patients, are a valuable starting point for devising important items. Two methods for gathering ideas for investigation are using focus groups and key informant interviews.[6] Focus groups are usually small collections of people who are allowed to informally discuss pertinent topics related to the investigational objective. Construction of these groups can be complex and important in shaping information gathered from these groups.[10] Key informant interviews differ from focus groups by consisting of smaller groups or even individuals who have unique knowledge and can be useful in exploring fields without much preexisting information.[6] The other major arm of item selection is clinical observation, which can range from anecdotal expert opinion to outcomes from more rigorous research. Patient-based perceptions and clinical observation may produce complementary items or, conversely, major divergent items.

Once items are agreed upon, the next step is to structure the language of the items that maximally obtains information from subjects. Expected level of knowledge, personal beliefs, and socioeconomic status will aid in producing questions easily understood and accepted by the subjects. Occasionally, additional education needs to be given to subjects with health surveys, especially when complicated concepts may be presented. Some additional education is often needed regardless of subject background; otherwise, areas of confusion may arise during pilot testing.

Table 1. Survey and Questionnaire Design

Objective	Determine Specific Need for Information
Data collection	Prospective
	Retrospective
	Experimental
Item selection	Expert opinion
	Focus group
	Key informant interviews
Question creation	Closed vs open
	Use of scales
	Order of questions to increase response
Sampling	Probability
	Nonprobability
Pilot testing	Performed informally
	Performed under expected conditions
Validity testing	Face validity
	Content validity
	Criterion validity
	Construct validity
Reliability testing	Test-retest reliability
	Intrarater reliability
	Interrater reliability
	Internal consistency
Administration	Face-to-face interview
	Telephone
	Mail
	Internet

Determine Question Structure and Order

Question construction has multiple parts. Questions can take two major forms: closed and open.[5] Both forms have strengths

and weaknesses. Closed questions can be either "yes/no" or multiple choice. This allows for increased ease of scoring and comparing results from questionnaires, which ultimately increases efficiency in reporting data within surveys.[9] The downside of closed questions is that potential answers may not be included, and they decrease the breadth of response and can take an unnatural form.[11] Conversely, open questions allow respondents to place fill-in responses that increase accuracy and individuality. Researchers can often gauge importance of a certain issue better, but scoring and comparison become very challenging.[12] In general, more thought needs to be put into the formulation of closed questions, whereas more energy is usually placed in interpretation of data with open questions.[2]

Both open and closed questions need to use simple, concrete, and nonconfusing language. Avoidance of biased phrases and words is paramount in avoiding bias within questions and response. Brevity decreases confusion; complex questions and answer formats decrease the yield of usable data.[13] Sentences should be complete and avoid 2-part questions that are ambiguous (example: "Do you have shortness of breath and nasal obstruction?").[14]

Questions also can be categorized by type of response, which include nominal, ordinal, or continuous. Nominal questions (eg, Question: what is your occupation? Answer: multiple choice of common occupations) should be both exclusive and inclusive. Exclusive means not having answers that overlap (ie, nurse and health care worker). Within reason, inclusive response selections to nominal questions should also be exhaustive. Ordinal questions take the form of rating or ranking an item. Scale ranges should be kept reasonable. A common ordinal question familiar to medical professionals is "rate your pain 1 to 10." Finally, continuous variable questions, which usually produce discrete data, can be used. An example of a continuous question would be "How many medical journals do you receive?"[2]

Following the design of individual items, the order of questions within a questionnaire then must be addressed. The introduction, including directions, should be simple and clear. Initial questions should be easy, close-ended, and attention grabbing.[15] Starting with an open-ended question often fails to capture attention and projects an image of difficulty to the responder. If the questionnaire is self-administered, it is preferable to keep the number of open-ended questions to a minimum. General questions should precede more specific questions, and questions with a time component should be place in chronological order.[9] Finally, demographic data should be placed at the end for 2 reasons, the first being a potential perception of intrusion and the second because they are generally easy to answer.[16]

Sampling

Interpretability and conclusions of results from a survey will strongly rely on who responds. The more representative a group of subjects is of a given population, the stronger and more applicable the findings will be. Thus, how sampling is performed plays a major role in study design. If a small population exists who are eligible for a given study (eg, a rare genetic disorder), then attempting to obtain responses from the whole population becomes possible; otherwise, sampling becomes necessary. Sampling falls into 2 broad categories, probability and nonprobability.[9]

In probability sampling, systematic approaches are used to decrease skewed groups. Probability sampling methods are the most powerful way to decrease bias in samples. Random techniques are probability approaches. Examples include simple random sampling, in which every member has an equal chance of being chosen (ie, a lottery), and stratified random sampling, in which subjects are placed into groups ahead of time according to a variable that strongly influences the outcome (eg, presentation with incomplete facial paresis vs complete facial paralysis). Randomization of each stratum occurs separately in stratified random sampling. A major advantage for using a stratified random sample is making sure a specific group is represented within the sample. Other nonrandom forms of probability sampling include systematic sampling where preset criteria are used to choose subjects (ie, using every nth subject) and cluster sampling that uses natural or preconceived groupings (ie, school districts).

Nonprobability sampling methods require less effort and cost in implementing. One of the most commonly used types of nonprobability sampling is convenience sampling, which relies on volunteers or easily obtained subjects, such as consecutive new patients. Another method includes using an index person or population for introduction to other individuals. A classic example is studying people engaged in either socially unacceptable practices or criminal activity where easy access to individuals is restricted. Utilization of quotas and usage of focus groups are also types of nonprobability sampling.[2,5]

Also important and related to sampling is calculating sample size. This is especially true when examining differences between 2 groups. Components needed to calculate sample size are alpha, beta (and its derivative, power), effect size, and estimate of deviation. The details of sample size calculation are beyond the scope of this article but can be found in a recent review.[17]

Administration

Many different forms of administration of surveys exist. Implementation can take place personally with an interviewer, be mailed, be conducted over the telephone, and increasingly be made available online. Although there is some commonality in administration of surveys, often emphasis changes depending on delivery methods. In general, self-administered instruments require simple and clear instructions.[15] If a scale is being used that scores the complete instrument, one poorly worded question can sidetrack the whole scale. Surveys that rely on interviews have a higher need for standardization of scoring by those recording responses.[18] Moreover, the delivery of questions also needs to be executed in a uniform manner.

Many of these issues are discovered during pilot testing of a new instrument. Pilot testing has 2 main functions—discovering flaws within the instrument and also examining

Table 2. Validity in Order of Power

Face validity	Face validity suggests the instrument appears to measure what it is supposed to measure. An example might be an eye-hand dexterity test to evaluate a component of surgical skill. This is the least powerful validity test.
Content validity	Content validity refers to the fact that the items make sense and comprehensively cover the issue. It requires that the universe of content items germane to the issue be included and that content unrelated to the issue be excluded. A panel of experts and several revisions are usually required. An example might be a series of questions regarding study habits when trying to improve resident selection.
Criteria-related validity	This test of validity compares the new "target test" against a "gold-standard" criterion. This test may be (1) concurrent or (2) predictive.
(1) Concurrent validity	Concurrent validity refers to the target test and the gold-standard criterion being conducted at the same time. An example might be evaluating a new auditory test in comparison with an auditory brainstem test to determine the new test's validity in detecting an eighth nerve dysfunction.
(2) Predictive validity	Predictive validity refers to how well the target test is able to predict the results from a gold-standard criterion obtained at some time in the future. An example might be a clinical tool focused on clinical signs, symptoms, and office tests that might be predictive of magnetic resonance imaging discovery of a solitary vestibular schwannoma.
Construct validity	A construct is a psychological, abstract concept that is difficult or impossible to measure. To determine if an instrument has construct validity, the instrument must have strong content validity relative to the construct to be tested and defined theoretical context. "An instrument is said to be a valid measure of construct when the measurements support these theoretical assumptions."[20] An example might be an instrument to be used in resident selection that might predict how well the resident will ultimately be in the future.

the reliability and validity of the questionnaire (see Validity and Reliability below). Ideally, pilot testing is performed on a separate group than the group used ultimately for data. Different phases of testing can take place, with early phases being more informal, and often consist of giving instruments to other professionals within a field. The advantage here is that problems with question design are discovered early.

Later phases of pilot testing usually are implemented in a manner closer to how it will eventually be administered and with subjects closer to the intended sample. During this phase, administration flaws and the ability to apply psychometrics to the instrument can occur. These later phases of pilot testing are crucial for eliminating variance produced by interviews. Pilot testing of new instruments often becomes the subject of research, and many choose to report these data prior to using the instrument in its intended purpose.[2,5,9]

Maximizing response rates is a major challenge of any survey. Higher response rates add credibility to the results. Fundamental to any study is the inference that the sample reflects the general nature of the universe (ie, how generalizable the results are to larger populations). No exact response rate can be considered sufficient. Lower response rates (below 50%) may be tolerable if the absolute number is still large (eg, 3000 responses after 20,000 contacts initially made). Conversely, if the initial target sample is small, response rates below 70% may be insufficient. If low response rates are present, the important question is, what would the nonresponders answer? If one sample population is studied, the generalizability of the results is primarily questioned. If a comparison study is being performed, nonresponders may seriously affect the baseline comparability of the groups. One area where designers have control in increasing response is being sure the sampled population is very interested in the issue being addressed. Another method of increasing responses is in designing simple and clear questions that follow a logical order in as brief as possible questionnaire. This decreases either not answering or answering in unintended formats, which cannot be recorded in an ideal manner. Assured anonymity increases individual participation. Offering details in how privacy/confidentiality issues are addressed within surveys will increase participation. If surveys are mailed or e-mailed, follow-up contact increases response.[15,19] An obvious way for increasing responses is giving incentives. Finally, field testing the questionnaire on smaller samples to determine the response rate and to determine the exact reasons for nonresponse are crucial for effective questionnaire design. This allows correction of fatal flaws and insights into the populations being questioned. Spending time and labor on the design of a good questionnaire drawing high responses is far better than trying to explain inadequate results at the end.

Validity and Reliability

The concepts of reliability and validity are psychometric measures that come from classic test theory.[6] Validity refers to the test measuring what it is intended to measure[20] (**Table 2**). For example, measuring from the bottom of the feet to the top of the head of a standing man is a valid measure of the man's height but only if the tool used to measure is reliable. Reliability means getting close to the same results each time the measurement is taken (**Table 3**). Thus, validity requires reliability.[20] However, a reliable instrument may not be valid. For example, a scale that reliably measures the man's weight is not a valid measure of the man's height. Also, the same ruler that reliably measured the man's height may also reliably measure the man's foot length, but measuring the foot would not be a valid measure of the man's height. Reliability and validity are usually examined during pilot administration of the questionnaire.

Table 3. Reliability

Test-retest reliability	Testing the same subjects twice, with an appropriate time interval between tests, and getting close to the same result for each subject
Intrarater reliability	Same rater testing the same subjects 2 or more times, with an appropriate time interval between tests, and getting close to the same results for each subject
Interrater reliability	Two or more raters testing the same subjects with the same instrument and getting close to the same results
Internal consistency (homogeneity)	"Internal consistency, or homogeneity, reflects the extent to which items measure various aspects of the same characteristic and nothing else."[20] A classic measure of this is Cronbach's coefficient alpha used with dichotomous or multiple-choice data. A high value of alpha is expected; however, if alpha significantly increases when an item is left out, that would suggest the item might not be homogeneous and could be removed.[6]

Management

Once responses are obtained, management of the data becomes key for successful utilization of questionnaires. Usually the data are compiled within a spreadsheet with use of a preset code that allows easy interpretation and manipulation of the data. Electronic responses (eg, from e-mail or specific Web sites) allow for direct placement of responses within such a system. Otherwise, data need to be placed manually. It is during this stage of survey data management that cleaning of data occurs, which means addressing missing data or incorrectly answered data.[5,9]

Various options exist regarding management of missing data, with the most conservative consisting of complete dismissal of the data and the most aggressive being filling data based on estimation. Determining the characteristics of nonresponders is important in deciding how to manage data. Several methods can be used to extrapolate the characteristics of nonresponders, including comparing late responders' answers and reasons (especially those who needed extra contact or incentive) to those of early responders, examining characteristics from other similar studies, or data from pilot tests of the questionnaire. If one can show that responders and nonresponders do not differ largely in other domains (ie, demographics or symptomatology), simply disregarding lack of response is an easy way to handle this situation. Another technique is weighting the data such that nonresponse is minimized.[9] Perhaps the most aggressive way of dealing with nonresponse is using imputation, in which responses are placed. This can be done by randomly assigning answers (which can save an entire scale if only a few responses are missing) or by looking at how similar responders answered questions and then placing the most common answer.[21,22] These techniques, if used, change the data and should not be employed often. A similar concept is how to deal with outlying data. Again, the most conservative way of dealing with these data is to keep them in the analysis. Others argue that minimally removing the most extreme outliers adds clarity.[2]

Biases

Bias can be introduced either from the designers or responders of a survey. Bias is insidious, and continuous vigilance is necessary to both recognize and minimize it. Bias produced by those constructing instruments and their implementation includes both question and questionnaire design. Biases in question design can be broadly broken down into problems with wording, incomplete data, use of faulty scales, leading questions, and inconsistency. Similarly, formatting, length of questionnaires, and flawed structure are general types of questionnaire design biases. These forms of bias are most readily controlled by project design. Often following sound construction of questions and questionnaires as described above and elsewhere will decrease these forms of systematic error.[23]

The other source of bias is from responders. Responders need to have the cognitive ability to read, interpret, and answer questions. Furthermore, responders' subconscious and conscious tendencies and cultural differences all produce bias. Knowing about these biases allows utilization of techniques to minimize them.[24]

Cognition and bias interface in several domains. Responders first need to understand the question. This interrelates to question design and understanding the study's population. Recall ability is another cognitive source of bias. Chronic issues, including those that fluctuate over time, have been shown to vary widely when compared with diaries over the same period.[25] Pain especially is reported to be difficult to assess at later time points.[26-28] Furthermore, people tend to underestimate common occurrences and overestimate rarer occurrences. Responders may display end-digit bias, where estimation of number of events ends either in "0" or "5."[6]

Subconscious and conscious interactions take place in responding to questions. Examples of subconscious effects include avoiding extreme answers (agree vs strongly agree—known as central tendency) and responding in a generally affirmative way, especially when asked about satisfaction. Conscious forms of bias include "faking good," when a subject wishes to be seen in a positive light, and "faking bad," when it is assumed reporting a worse situation is to the benefit of the subject. Often these forms of bias arise surrounding socially unacceptable circumstances (eg, sexually transmitted diseases or smoking during pregnancy).[23] Anonymity helps to decrease these types of bias.

Special situations that inject bias often in surveys are cultural differences and proxy reporting. Bias due to culture can occur both from interpretation of questions as well as in the response. Considerations should be made if an expected large proportion of a sampling is of a similar culture, which may handle certain

Table 4. Key Points for a Successful Survey and Questionnaire

Highly interested study sample
Sharply defined objectives
Short, clearly written, nonambiguous questions
Logically arranged questions
Brief questionnaire
Guaranteed anonymity
Easy response method
Opportunity to recontact subjects
Field-tested instrument and administration
Proven reliability
Proven validity
Incentives

topics different from others within the sample. Questions should be made understandable to all groups.[6] Another source of bias occurs when someone else is answering for a subject. This often occurs when people are either physically or mentally unable to give responses for themselves. Proxies tend to be more reliable when stating objective answers to questions (ie, how many cigarettes does she smoke?) as opposed to internal/emotional questions (ie, what is his quality of life?).[29,30]

Conclusions

Surveys and questionnaires are powerful research tools. The correct construction, implementation, and management of these research tools are critical in creating meaningful data. Bias must be minimized. Moreover, understanding of these principles increases one's ability to interpret and use information from these sources effectively in patient care. **Table 4** itemizes the key issues to be addressed in reading or developing a quality survey and questionnaire.

Author Contributions

Eric L. Slattery, substantial contributions to conception and design, acquisition of data or analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and final approval of the version to be published; **Courtney C. J. Voelker**, substantial contributions to conception and design, acquisition of data or analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and final approval of the version to be published; **Brian Nussenbaum**, substantial contributions to conception and design, acquisition of data or analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and final approval of the version to be published; **Jason T. Rich**, substantial contributions to conception and design, acquisition of data or analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and final approval of the version to be published; **Randal C. Paniello**, substantial contributions to conception and design, acquisition of data or analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and final approval of the version to be published; **J. Gail Neely**, substantial contributions to conception and design, acquisition of data or analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and final approval of the version to be published.

Disclosures

Competing interests: None.

Sponsorships: None.

Funding source: None.

References

1. Piccirillo JF, Merritt MGJ, Richards ML. Psychometric and clinimetric validity of the 20-item Sino-Nasal Outcome Test (SNOT-20). *Otolaryngol Head Neck Surg*. 2002;126:41-47.
2. Fink A. *The Survey Kit*. 2nd ed. Thousand Oaks, CA: Sage; 2003.
3. Groves RM. *Survey Methodology*. 2nd ed. Hoboken, NJ: John Wiley; 2009.
4. Rea LM, Parker RA. *Designing and Conducting Survey Research: A Comprehensive Guide*. 3rd ed. San Francisco: Jossey-Bass; 2005.
5. Fink A. *How to Conduct Surveys: A Step-by-Step Guide*. 4th ed. Thousand Oaks, CA: Sage; 2009.
6. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. New York: Oxford University Press; 2008.
7. Rust J, Golombok S. *Modern Psychometrics: The Science of Psychological Assessment*. 3rd ed. Hove, UK: Routledge; 2009.
8. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. New York: McGraw-Hill; 1994.
9. Aday LA, Cornelius LJ. *Designing and Conducting Health Surveys: A Comprehensive Guide*. 3rd ed. San Francisco: Jossey-Bass; 2006.
10. Vogt DS, King DW, King LA. Focus groups in psychological assessment: enhancing content validity by consulting members of the target population. *Psychol Assess*. 2004;16:231-243.
11. Smith TW. The art of asking questions, 1936-1985. *Public Opin Q*. 1987;51:S95-S108.
12. Geer JG. Do open-ended questions measure "salient" issues? *Public Opin Q*. 1991;55:360-370.
13. Payne SLB. *The Art of Asking Questions*. Princeton, NJ: Princeton University Press; 1951.
14. Fowler FJ. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage; 1995.
15. Dillman DA. *Mail and Internet Surveys: The Tailored Design Method*. 2nd ed. Hoboken, NJ: John Wiley; 2007.
16. Tourangeau R, Rips LJ, Rasinski KA. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press; 2000.
17. Neely J, Karni R, Engel S, Fraley PL, Nussenbaum B, Paniello RC. Practical guides to understanding sample size and minimal clinically important difference (MCID). *Otolaryngol Head Neck Surg*. 2007;136:14-18.
18. Cannell CF, Oksenberg L, Converse JM, et al. *Experiments in Interviewing Techniques: Field Experiments in Health Reporting, 1971-1977*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan; 1979.
19. Fowler FJ. *Survey Research Methods*. 4th ed. Thousand Oaks, CA: Sage; 2009.
20. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 3rd ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2009.

21. Brand JPL, van Buuren S, Groothuis-Oudshoorn K, Gelsema ES. A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*. 2003;57:36-45.
22. Groves RM. *Survey Nonresponse*. New York: John Wiley; 2002.
23. Choi BC, Pak AW. A catalog of biases in questionnaires. *Prev Chronic Dis*. 2005;2:A13.
24. Jabine TB; US National Research Council; Committee on National Statistics. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines: Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology*. Washington, DC: National Academy Press; 1984.
25. Means B; US National Center for Health Statistics. *Autobiographical Memory for Health-Related Events*. Hyattsville, MD: US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics; 1989.
26. Salovey P, Smith AE, Turk DC, Jobe JB, Willis GB. The accuracy of memory for pain: not so bad most of the time. *APS J*. 1993;2:181-191.
27. Stone AA, Broderick JE, Kaell AT, DelesPaul PA, Porter LE. Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritics? *J Pain*. 2000;1:212-217.
28. Stone AA, Broderick JE, Shiffman SS, Schwartz JE. Understanding recall of weekly pain from a momentary assessment perspective: absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain*. 2004;107:61-69.
29. Todorov A, Kirchner C. Bias in proxies' reports of disability: data from the National Health Interview Survey on disability. *Am J Public Health*. 2000;90:1248-1253.
30. Yip JY, Wilber KH, Myrtle RC, Grazman DN. Comparison of older adult subject and proxy responses on the SF-36 health-related quality of life instrument. *Aging Ment Health*. 2001;5:136-142.